2D and 3D CT Radiomics Features Prognostic Performance Comparison in Non-Small Cell Lung Cancer

Chen Shen[*,†,1], Zhenyu Liu[†,1], Min Guan[‡,1], Jiangdian Song[†,§], Yucheng Lian[†], Shuo Wang[†], Zhenchao Tang[†,¶], Di Dong[¶], Lingfei Kong[‡], Meiyun Wang[‡], Dapeng Shi[‡] and Jie Tian[*,†,#]

[*]School of Life Science and Technology, Xidian University, Xi'an, Shaanxi, 710126, China; [†]CAS Key Laboratory of Molecular Imaging, Institute of Automation, Beijing, 100190, China; [‡]Department of Radiology, Henan Provincial People's Hospital & the People's Hospital of Zhengzhou University, Zhengzhou, Henan, 450003, China; [§]Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang, 110819, China; [¶]School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, Shandong Province, 264209, China; [#]University of Chinese Academy of Sciences, Beijing, 100080, China

Abstract

OBJECTIVE: To compare 2D and 3D radiomics features prognostic performance differences in CT images of non-small cell lung cancer (NSCLC). *METHOD:* We enrolled 588 NSCLC patients from three independent cohorts. Two sets of 463 patients from two different institutes were used as the training cohort. The remaining cohort with 125 patients was set as the validation cohort. A total of 1014 radiomics features (507 2D features and 507 3D features correspondingly) were assessed. Based on the dichotomized survival data, 2D and 3D radiomics indicators were calculated for each patient by trained classifiers. We used the area under the receiver operating characteristic curve (AUC) to assess the prediction performance of trained classifiers (the support vector machine and logistic regression). Kaplan–Meier and Cox hazard survival analyses were also employed. Harrell's concordance index (C-Index) and Akaike's information criteria (AIC) were applied to assess the trained models. *RESULTS:* Radiomics indicators were built and compared by AUCs. In the training cohort, 2D_AUC = 0.653, 3D_AUC = 0.671. In the validation cohort, 2D_AUC = 0.755, 3D_AUC = 0.663. Both 2D and 3D trained indicators achieved significant results ($P < .05$) in the Kaplan-Meier analysis and Cox regression. In the validation cohort, 2D Cox model had a C-Index = 0.683 and AIC = 789.047; 3D Cox model obtained a C-Index = 0.632 and AIC = 799.409. *CONCLUSION:* Both 2D and 3D CT radiomics features have a certain prognostic ability in NSCLC, but 2D features showed better performance in our tests. Considering the cost of the radiomics features calculation, 2D features are more recommended for use in the current study.

Translational Oncology (2017) **10**, 886–894

Introduction

Lung cancer is one of the most common and deadly cancers in the world, most of which are non-small cell lung cancer (NSCLC) (85% to 90%) [1]. Current NSCLC guidelines suggest survival data having a correlation with staging [2]. Both NSCLC patients' prognosis and medical resources can benefit "personal treatment" from properly predicted factors of refined subgroups [3]. Computed tomography (CT) is the common tool for NSCLC diagnosis [4]. Many prior

http://dx.doi.org/10.1016/j.tranon.2017.08.007

studies have worked on investigating multiple CT imaging factors and indicators to improve the prediction, such as using image texture analysis to measure the tumor heterogeneity and permeability [5].

"Radiomics" was proposed and introduced into clinical oncology studies in the past 5 years. It converts medical images into numerical features, and uses data-mining algorithms or statistical tools for further analysis. By building appropriate models with refined features, it has shown successful assessment and prediction abilities and applied them in various challenging clinical tasks. Huang et al. found that the radiomics features could be a potential biomarker for prognostic prediction of the free survival in NSCLC [6]. Coroller et al. demonstrated that the radiomics features can predict distant metastasis in lung adenocarcinoma [7]. Aerts et al. have shown the association of radiomics features' prognostic power and NSCLC gene expression [8].

Radiomics research requires a huge amount of quantitative imaging data [9]. Certain radiomics-based tools and system have been developed and applied recently [10]. These tools could help researchers in deriving implicit image features. To pick stable and high-performance radiomics features is crucial for further study. Balagurunathan et al. discussed the reproducibility of CT image features in the NSCLC survival analysis [11,12]. However, when extracting CT radiomics features, there is a basic trade-off. Tumor lesions in CT images are expressed in multiple layers, thus one can calculate features with the whole layers in 3D, or just calculate one typical 2D layer's features (such as the layer with the largest lesion cross-section). 2D features are easier to obtain with less labor consumption, lower complexity and faster calculation. Intuitively, 3D features are opposite to 2D's but might carry more information about the tumor. Both 2D and 3D features have been employed in past studies [6,8,13,14]. The performance differences between 2D and 3D radiomics features have not been discussed yet. This issue is essential for further popularization of radiomics in clinical research.

In this paper, our hypothesis is that 3D radiomics features have higher performance than 2D. To test this assumption, we assessed a total of 1014 radiomics features in three independent cohorts involving 588 NSCLC patients. We handled 2D and 3D features groups (507 per group) into a series of parallel experiments correspondingly. The selected features and statistical tools are commonly used and reported in prior papers [6–8]. Therefore, our investigation could refer to further radiomics-based clinical research and diagnosis system building.

Materials and Methods
This retrospective study was approved by the institutional review board of Henan Provincial Hospital, which waived the requirement for patients' informed consent. Medical record review was performed in accordance with the institutional ethics review board guidelines. For using public data, we follow the citations and data usage policy from "the cancer imaging archive" (TCIA) public access, which is maintaining and operating the largest global public cancer image-sharing platform [15]. The research sequence is illustrated in Figure 1. Details could be checked in the following sections.

Patients
In the present study, we collected 588 NSCLC patients' pre-treatment CT images from three cohorts. The inclusion criteria of patients are: (a) patients who have no treatment before the image taking; (b) patients who had pathologically confirmed grading results;

(c) aged over 18 years; and (d) CT images' slice thickness less than 5 mm.

Images of the training cohort were downloaded from the TCIA website (http://www.cancerimagingarchive.net/). One dataset was named as "NSCLC-Radiomics" provided by the Department of Radiation Oncology (MAASTRO), Maastricht University, The Netherlands. There are 422 patients received an FDG PET-CT scan for radiotherapy treatment planning in a radiotherapy position on a dedicated PET-CT simulator with both arms above the head. For the FDG PET-CT scans, a Siemens Biograph (SOMATOM Sensation-16 with an ECAT ACCEL PET scanner) was used. An intravenous injection (weight * 4 + 20) MBq FDG (Tyco Health Care, Amsterdam, The Netherlands) was followed by 10 ml of physiological saline. After a 45-min uptake period, during which the patient was encouraged to rest, PET and CT images were acquired. A spiral CT (3 mm slice thickness) with or without intravenous contrast was performed covering the complete thoracic region. We removed 15 cases from the original dataset for image reading issues, which included all suspicious cases without intravenous contrast. The other dataset of "TCIA" was named as a "LungCT-Diagnosis" provided by the Moffitt Cancer Center with 61 patients. Specifically, the slice thickness of CT images was between 2.5 mm to 6 mm. We also removed five cases from the original dataset for tumor segmentation issues and the slice thickness was over 5 mm. The protocol was approved by the Institutional Review Board (IRB). Acquisition details could be looked up in their published papers [8,16].

There are 125 images from the validation cohort were acquired from the Department of Radiology at the Henan Provincial People's Hospital during 2012 to 2015. Chest contrast-enhanced CT was performed on every patient using one of the two multi-detector row CT (MDCT) systems (Philips Brilliance 16 slices CT, Phillips Medical System, the Netherlands or 64-slice LightSpeed VCT, GE Medical systems, Milwaukee, USA), with the following acquisition parameters: 120 kV; 160 mAs; 0.5 or 0.4 s rotation time; detector collimation: 16×1.25 mm or 64×0.625 mm; field of view, 350×350 mm; matrix, 512×512. After routine non-enhanced CT, contrast-enhanced CT was performed after a 40s delay following intravenous administration of 85 ml of iodinated contrast material (Ultravist 370, Bayer Schering Pharma, Berlin, Germany) at a rate of 2.5–3.0 ml/s with a pump injector (Ulrich CT Plus 150, Ulrich Medical, Ulm, Germany). The CT image was reconstructed with a standard kernel. The slice thickness of the CT images was in the range between 0.625 mm and 1.25 mm. These CT images were retrieved from the picture archiving and communication system (PACS) (Huahai, China). We only collected contrast-enhanced CT images for the further analysis.

These three cohorts were independent and collected from different centers. The survival end events were defined as the date of death. The demographic statistics are listed in Table 1.

Tumor Segmentation
We performed the segmentation of lung lesions semi-automatically. All images were read and processed in the raw DICOM format. Radiologists with over 10 years of experience examined each layer of the patients' CT data, and identified a proper point inside of the lesions. This point would become a "seed point". Afterwards, we used a "toboggan based" growing automatic segmentation approach to deal with those seed points [17]. Toboggan based growing algorithm is an automatic method for lung lesion segmentation. The original seed point
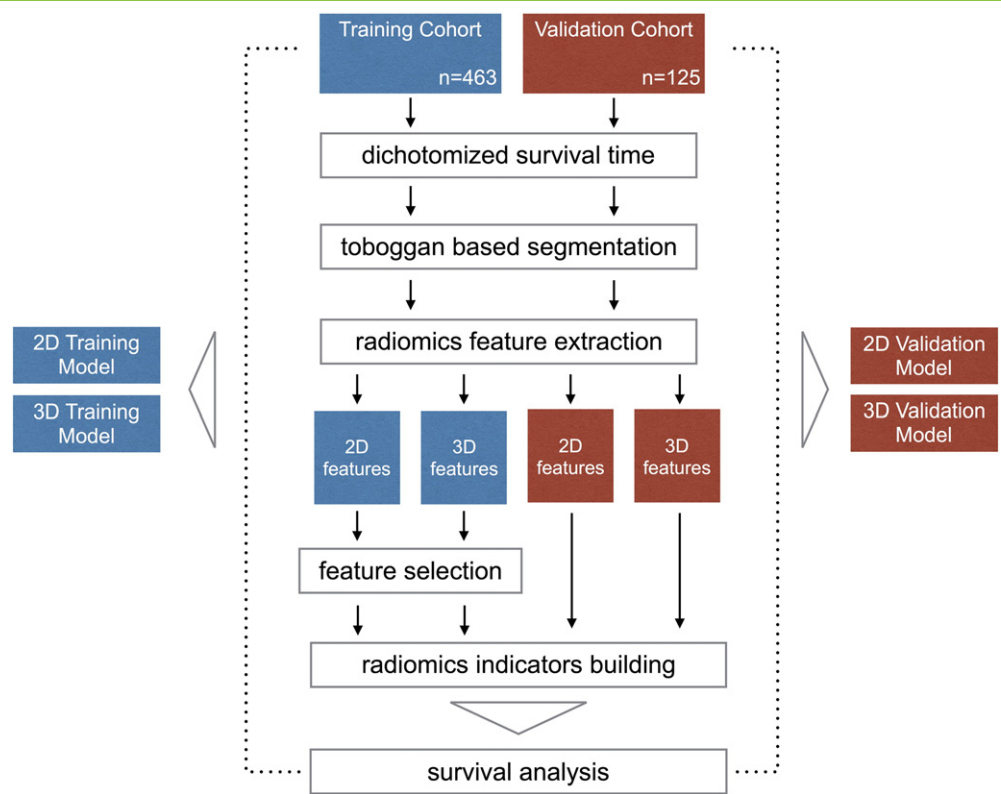
Figure 1. The data flow sequence in this research. All patients' survival data were dichotomized by the cut-off of 2 years (1, larger than 2 years; 0, less than 2 years). Then, tumors' contours were segmented by our automatic algorithm. We extracted the 2D and 3D radiomics features based on those segmentations. In the training cohort, we selected the extracted features depending on some rules. Based on the selected features, we built and validated the radiomics indicators. Finally, the survival analysis was correlated with the survival time and radiomics indicators. We then compared the 2D and 3D model in both the training and validation cohorts.

could be obtained automatically by the improved toboggan algorithm. Twenty-six neighborhoods of each voxel were calculated to determine whether it could be included in the lung lesion region. Lesion boundary was automatically determined by the result of the improved toboggan algorithm. The accuracy of this method is 81% on average compared with manual segmentation. Particularly, the segmentation algorithm performed well for ground-glass nodules (86% compared to the radiologists). This algorithm has been programmed as an add-on in the Medical Imaging Toolkit (MITK), which is a C++ library for integrated

medical image processing and is developed by the Institute of Automation, Chinese Academy of Sciences [18]. Finally, 3D regions of tumors would be generated from "seed points" and highlighted. Radiologists checked these segmentations until they were satisfied. Since this segmentation method is based on the calculation of raw images, it can minimize the experience of deviation between different radiologists in order to improve the stability of the segmentation results, and can improve this research efficiency. Figure 2 shows radiologists working on this segmentation program and an example of the segmentation result.
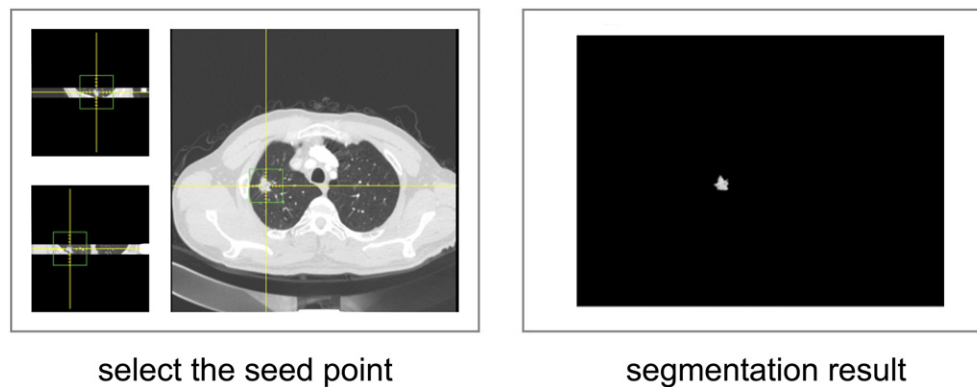
Radiomics Feature Extraction

We implemented this calculation procedure through our home-made Matlab scripts (Matlab 2014b; Mathworks, Natick, MA, USA). Image resampling is crucial preprocessing for 3D features extraction, since the slide thickness of the data sets is different. We used Matlab functions ("resample" and "reshape") for down sampling or up sampling that allows all of the CT images to be adjusted to 3 mm slice spacing. Feature extraction was based on the segmentation results from the previous section. For the 3D features group, we calculated 507 features in the labeled pixels inside the lesion contours. In contrast, for the 2D features group, we calculated 507 features in the labeled pixels for the layer that includes the maximum cross-section of the lesion. As designed, the 3D features group and the 2D features group should be one-to-one correspondingly. Both 2D and 3D features groups involve certain categories: first-order histogram statistics with 14 features, Gray-Level Co-occurrence Matrix

Table 1. Characteristics of Patients in the Training and Validation Cohorts

Characteristics		Training Cohort	Validation Cohort	P-Value
Number of Patients		463	125	
Gender				.644
	Male	310	83	
	Female	173	42	
Age (years)				<.001 ,†
	Range	43–91	39–83	
	Median	68	63	
Survival				.135
	Median (days)	462	482	
	No. >2 years	131	44	
Overall stage				.572
	I	110	33	
	II	55	11	
	III and IV	298	81	

Note: p-values are results of χ^2 test for categorized variables and t test for continuous variables. $P < .05$.

† Since the age information for "LungCT-Diagnosis" is missing, we only compared the other two datasets.

select the seed point   segmentation result

Figure 2. Screenshots of experienced radiologists working on the segmentation program, and the tumor segmentation result.

(GLCM) with 12 features, Gray-Level Run-length Matrix (GLRL), and Fractal Dimension with 13 features [19–21]. On the other hand, totally 12 image filters with multiple scales were also employed: Gaussian filter, Laplacian filter and wavelet. Hence, we finally incorporated the combination of categories and multiple filters into the radiomics feature set. Details of radiomics features are listed in the Appendix A1.

Feature Selection

The statistical analysis was performed in R software (version 3.3.0; http://www.Rproject.org). The used R packages in this paper are listed in the Appendix. The reported statistical significance levels were all two-sided at $P = .05$. Our feature selection strategy was based on the features' stability and the classification performance of the prognostics. Considering the generalization of the conclusion among multiple data sources, the Kruskal-Wallis test (with random features grouping, bootstrap for 1000 times) was adopted among features in the training cohort to test each feature's stability. If the P-value after the Kruskal Wallis test is larger than the significant threshold, this indicates that this feature's distribution had no difference between cohorts, hence the selected features were "stable features". After that, we correlated the features with patients' survival data. The univariate Cox regression model was employed to achieve each feature's Harrell's concordance index (C-Index), which represented the features' classification performance to a certain extent [22]. Features with potential prognostic power would have higher C-Indices. Finally, we selected both stable and potential prognostic features to construct the 2D group's and 3D group's indicators of classification.

Statistical Tools

Selected features would be integrated as one indicator for the prognostic prediction. In order to build this predictor, we introduced the logistic regression as the classifier. Ten-fold cross-validation was used for parameters tuning in the training cohort. The employed classifiers were mainly applied to the binary classification issues. All censored continuous survival data were dichotomized by a cutoff of 2 years. Hence, we temporally defined survival labels of the patients: "1" represents those who live longer than the cutoff time and are labeled as "0". The cutoff time of 2 years referred to other studies whose prediction models used this survival cutoff [23–26]. It could be considered as a pertinent median survival time of NSCLC patients. If a patient was labeled as "1," this means it was in the high-risk group, otherwise it was in the low-risk group. The radiomics indicators were

calculated for each patient by trained classifiers. We used the area under the receiver operating characteristic (ROC) curve (AUC) to assess the prediction performance of the classifiers, and a t test was employed to assess the significant differences between the results. Univariate analysis (two-sample t test) was used to evaluate the rationality of the single selected feature. Next, the Kaplan–Meier analysis was used to associate radiomics indicators with the survival information in the validation cohort. The computed indicators were utilized for splitting the survival curves. We then used the log-rank test to assess significant differences between the two survival curves. The computed radiomics indicators were also associated with survival time and evaluated by the Cox hazard regression model. C-Index and Akaike's information criteria (AIC) were used to assessed the building of the Cox models [13,27]. Wilcoxon test was used to assess the significance of the calculated C-Indices of models by the bootstrap approach for 1000 times.

Control Experiments

We designed two control experiments to improve the comprehensiveness of our conclusion. We firstly removed patients with survival time in the interval of 1.5 to 2 years in the both cohorts and repeated the statistical analysis process. We hope it can amplify the difference between long and short survivals for better observations. Secondly, we selected the best 2D features and evaluated the performance of the corresponding features with 3D features. It can enhance the difference between 2D and 3D groups of features. We also repeated the same analysis process as above. These two control experiment were described in Appendix A3.

Results

Selected Features for the Radiomics Signature

Based on the Kruskal Wallis test, a total of 57 2D features and 47 3D features were filtered out. They had a similar distribution among the patient cohorts, therefore they could be considered as stable features. There were 15 corresponding features between the 2D and 3D groups in these stable features. Depending on the C-Index of the univariate Cox regression of those selected features, we refined eight features of each group to build the radiomics signature. These refined features with their C-Indices are plotted in Figure 3. Higher C-indices normally indicated higher prognostic performance of the features. The two-sample t test showed the selected features' C-indices (demonstrated in Figure 3) between the 2D and 3D groups that were
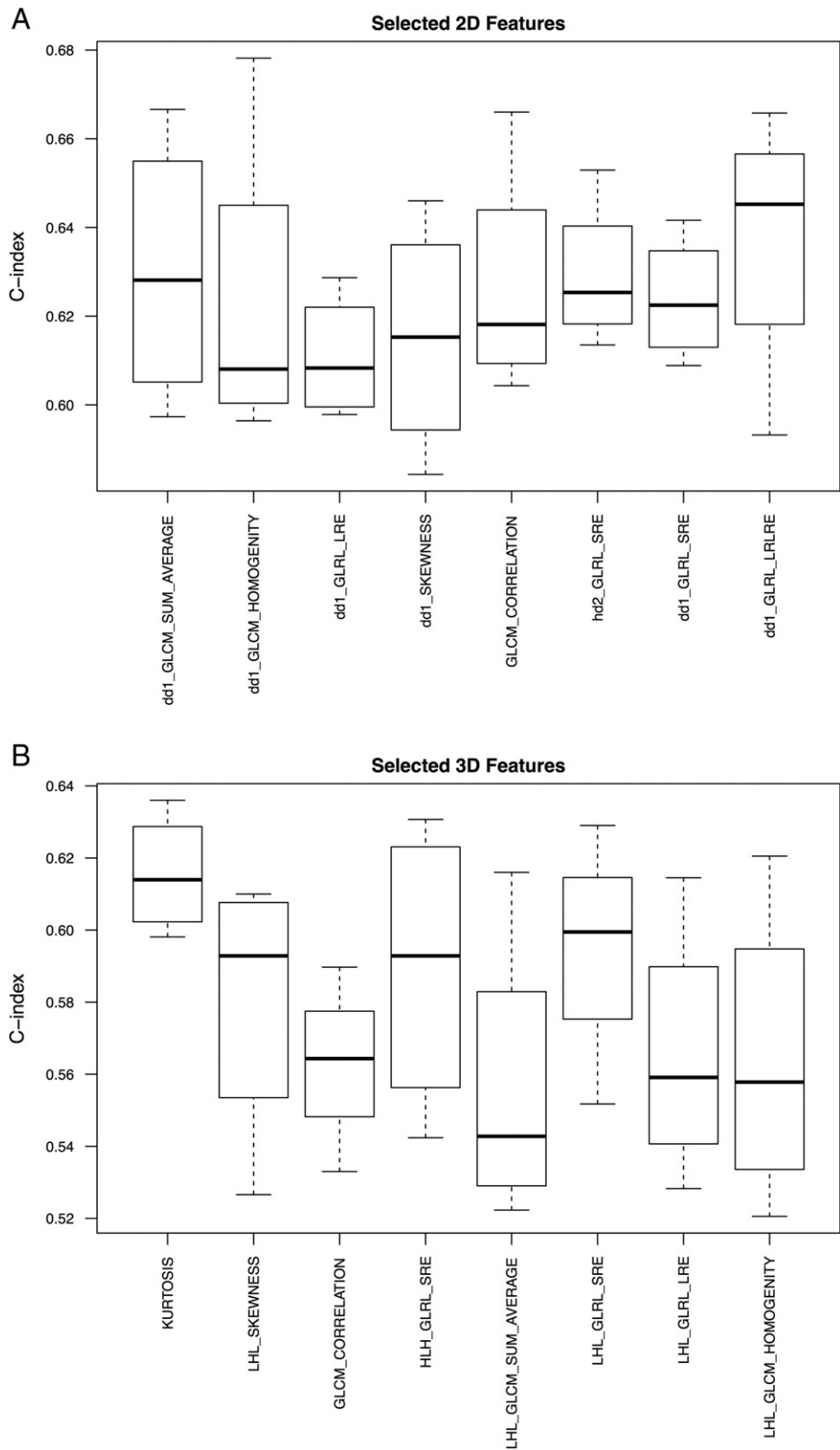
Figure 3. The refined radiomics features with their C-indices. (A) 2D features group; (B) 3D features group.

significantly different ($P < .001$). We also performed a univariate analysis on the selected features correlated with the dichotomized survival time. We compared each feature's difference between the high-risk group (labeled as "1") and the low-risk group (labeled as "0") in both the training and validation cohorts with the t test. Table 2 lists the p-values of each selected feature in the cohorts.

Survival Indicators Building

We implemented the logistical regression model depending on the dichotomized survival data and the 16 selected radiomics features (8 features for each group). We assessed the predicted performance of the logistic classifier by introducing ROC in both training and validation phases. Figure 4 depicts their ROCs for the 2D and 3D

Table 2. *t* Test Results of the Feature Comparison Between High-Risk and Low-Risk Groups

2D Selected Features	P-Value		3D Selected Features	P-Value	
	Training	Validation		Training	Validation
dd1_SKEWNESS	0.445	0.056	LHL_SKEWNESS	0.137	0.599
GLCM_CORRELATION	0134	0.036	GLCM_CORRELATION	0.700	0.004
dd2_GLRL_LRE	0.093	0.036	LHL_GLRL_LRE	0.138	0.046
dd1_GLCM_SUM_AVERAGE	0.025	0.031	LHL_GLCM_SUM_AVERAGE	0.026	0.895
dd1_GLCM_HOMOGENEITY	0.018	0.038	LHL_GLCM_HOMOGENEITY	0.063	0.048
hd2_GLRL_SRE	0.001	0.093	LHL_GLRL_SRE	0.193	0.744
dd1_GLRL_SRE	0.181	0.027	HLH_GLRL_SRE	0.372	0.672
dd1_GLRL_LRLGE	0.044	0.003	KURTOSIS	0.002	0.040

Note:
  P < .05.
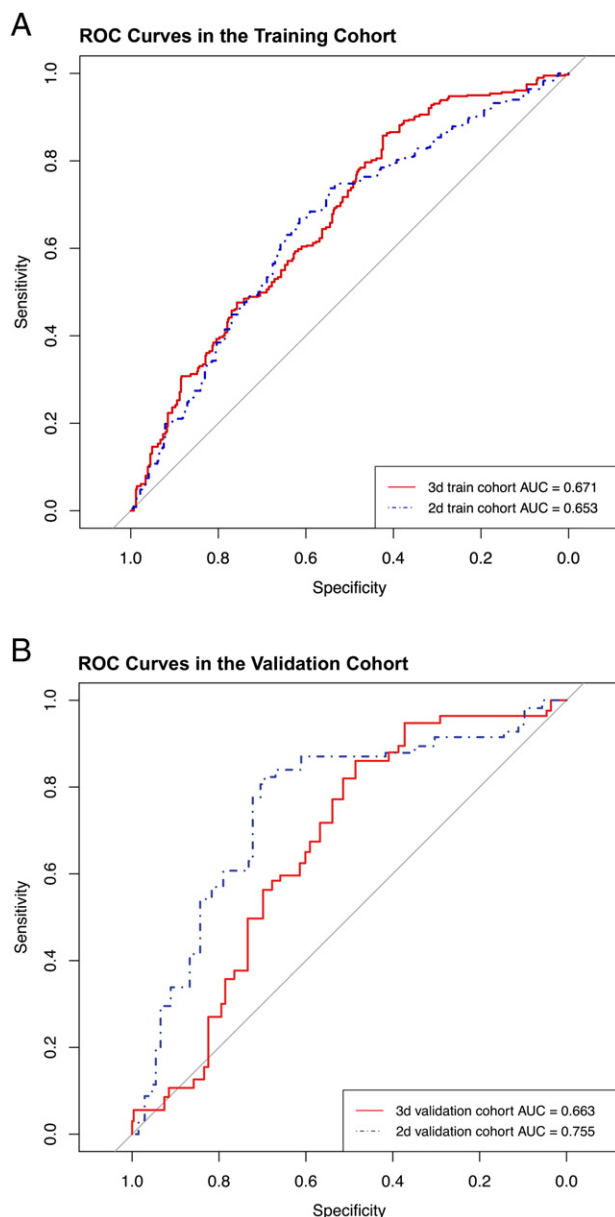
Figure 4. AUCs of radiomics indicators in both training and validation cohorts. (A) The model comparison in the training cohort: 2d_train_auc = 0.653, 3d_train_auc = 0.671; (B) The model comparison in the validation cohort: 2d_validation_auc = 0.755, 3d_ validation_auc = 0.663.

groups' comparison. For 2D and 3D indicators in the training cohort: 2D_AUC = auc = 0.653, 0.629 to 0.677 95% CI; 3D_AUC = 0.671, 0.647 to 0.694 95% CI. For indicators in the validation cohort: 2D_AUC = 0.755, 0.732 to 0.777 95% CI; 3D_AUC = 0.663, 0.640 to 0.687 95% CI.

Consequently, we achieved binary indicators of each patient in both the training and validation cohorts. Those indicators indicated whether patients were in the high-risk group or the low-risk group.

Survival Analysis

Classified binary indicators were associated with censored continuous survival data in the validation cohort. According to four categorized indicators, Figure 5 shows Kaplan Meier curves with significances of the log-rank test. The log-rank test results with the curves show that our radiomics indicator successfully divided the patients into high-risk and low-risk groups. Cox hazard regression was also applied to assess those indicators. Table 3 lists the analysis results of the hazard ratio (HR). Multivariable Cox regression results of the selected features are listed in the Appendix. In the validation cohort, the C-Index of the 2D model is 0.683 (0.651 to 0.716, 95% CI; *P* < .001, Wilcoxon test). The C-Index of the 3D model is 0.632 (0.600 to 0.669, 95% CI; *P* = .001, Wilcoxon test). The AIC of the 2D model is 789.047, and the AIC of the 3D model is 799.409. We have also achieved significant results of additional experiments. Details could be checked in the Appendix A4.

Discussion

In the present study, we compared the 2D and 3D radiomics features in NSCLC patients. We extracted a large number of quantitative features from 588 NSCLC patients' CT images from three different sources. We screened out 507 corresponding 2D and 3D features (1014 in total) of all cases in two independent cohorts. We hope the scale of our investigation can guarantee derived conclusions that are comprehensive and reliable. To define the features' performance in our study, we mainly assessed their stability and prognostic power. They were involved in a series of experiments and statistical tests. We would like to discuss those three properties below.

Because NSCLC lesions are mostly solid tumors, the performance of NSCLC radiomics features is highly dependent on segmentation. Numerous previous studies were based on manual segmentation, but the differences in segmentation outcomes among different radiologists must be considered. One article discussed radiomics feature differences between manual and automatic segmentation [11]. In a sense, the utilized semi-automatic segmentation method in our study
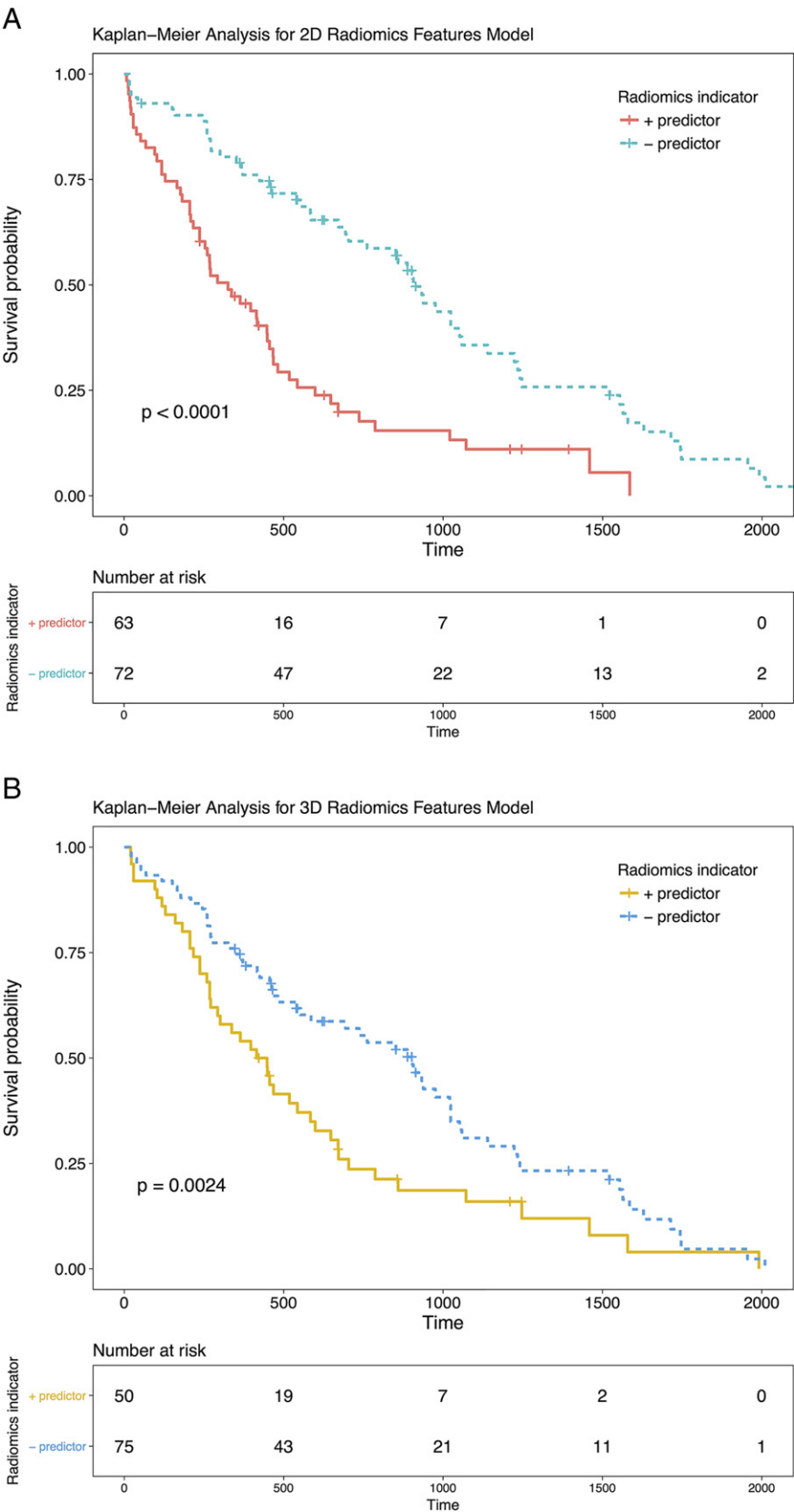
Figure 5. Kaplan–Meier analysis of the radiomics feature based indicators that split high risk and low risk groups in the validation cohort. (A) 2D model, *P* < .001, log-rank test (B) 3D model, *P* = .002, log-rank test.

could eliminate the experience differences from manual segmentation of radiologists. Under certain supervision and inspection, it takes an upfront guarantee to the features' stability before the feature extraction phase. Based on the Kruskal Wallis test results, we selected 57 3D features and 47 2D features correspondingly. In terms of the quantity and reduction process, 2D and 3D feature groups are fairly

Table 3. Risk of Radiomics Indicators

	HR	P-Value	95% CI for HR	
			Lower	Upper
2d_train	0.711	0.003	0.558	0.905
3d_train	0.692	0.001	0.546	0.877
2d_validation	0.421	<0.001	0.274	0.646
3d_validation	0.591	0.007	0.388	0.898

Note:
$P < .05$.

close. In other words, a similar number of stable features were found in the two groups. The selected features also tend to come from the same family (features with the same image processes or with similar wavelet filters). It could prove that the 2D and 3D feature groups are similar from a stability perspective.

We compared the selected features' prognostic power by a univariate analysis on features and multivariable regression models of the survival analysis. Most of our selected features belong to the same family, which includes the gray level texture features under the wavelet filter. We can point out that this feature family is very stable and represents the texture of the tumors significantly. As is known, the texture of the tumors has a deep association with tumor heterogeneity, as well as with the patients' prognosis [5,8,16]. The t test results for the C-Indices of the univariate Cox regression suggested that the features have differences in prognostic prediction, and 2D features seemed better (higher C-Indices in Figure 3). Table 2 listed the results of the univariate analysis of the features corresponding to the dichotomized survival time. We found multiple 2D and 3D features that were significantly different between the high-risk and low-risk groups. This also proved the rationality of our feature selection strategy, which screened out the stable and predictable features. The prognostic indicators were constructed by certain stable and predictable features with multivariable logistic regression. Figure 4A showed that a 3D group's AUC was slightly larger than the 2D's in the training cohort, but Figure 4B shows that the 2D's AUC is better in the validation cohort. The AUC results are reasonable and similar to a previous study [23]. Therefore, for the integrated use of these selected features, the 3D and 2D groups performed closely.

We employed the trained prognostic indicators and the survival data for the survival analysis. The training cohort and the validation cohort are independent. It would guarantee the reliability and credit of our conclusions. Figure 5 and Table 3 demonstrated that both 2D and 3D indicators achieved significant results in the Kaplan–Meier analysis and Cox regression. C-Index of the 2D model was slightly higher than the 3D model, but they all met the credible threshold. The AIC results of the 2D and 3D model were also similar. Referring to some former papers [7,8], we believe our created models are workable and meaningful.

In summary, 2D groups performed slightly better than 3D ones, since 2D had higher C-Indices in the selected features and a higher C-Index in the model comparison. In control experiments, we also achieved the similar results (details could be checked in Appendix A3). Due to this result, we believe that 2D features had better generalization ability. The reason for this result might be that the resolutions of the CT images were not consistent. The transverse plane's resolutions were different. This problem is difficult to avoid in multi-center studies and retrospective studies. Although we did the

resample processing before the radiomics feature extraction, it may still lead to errors. It is a slight deviation from the original definition of the 3D feature calculation, although we assumed 3D features might carry more dimensional information. This phenomenon is probably more pronounced in MRI or PET modalities, since their images are usually in thicker layers. On the other hand, it shows that 2D features could be sufficient for certain tasks, but 3D features are more time-consuming and require heavy-load computation. Along with the spread of new CT techniques, such as dual source CTs and further increase of computation power, these issues may be solved. Therefore, in the short-term future, 2D radiomics features are more recommended for clinical research and radiomics approaches based on software development. Along with the spread of new CT techniques, including dual source CTs and a further increase in computation power, this issue may need to be reopened. Several articles also discussed the repeatability, stability and classification issues of radiomics features [12,16]. We reproduced an elemental process of the radiomics research and achieved reasonable results. As more of the features were extracted in our multi-center source designed tests, this made our conclusions universal and more credible in general.

However, our prognostic indicators and models could be improved in many ways, such as introducing other feature selection frameworks. Another weakness of our study is that the 2-year cutoff is not equal to the cohorts' median survival time. It results in an imbalance between positive and negative dichotomized samples during classification tasks. This problem can be solved by expanding the database and choosing a proper cutoff [28]. In addition, we only implemented the radiomics approach of the NSCLC survival analysis. Except for the shown tasks, radiomics can be used for many other diseases and modalities. Consequently, our conclusion should be conservatively generalized.

Conclusion
The present study was aimed to compare the prognostic performance difference between 2D and 3D radiomics features in NSCLC. Both 2D and 3D features have a certain prognostic ability, but 2D features performed slightly better and were easy to achieve. It is recommended to choose 2D features in practical research.

Appendix A. Supplementary Data

References
[1] Siegel RL, Miller KD, and Jemal A (2016). Cancer statistics, 2016. *CA Cancer J Clin* **66**(1), 7–30.
[2] Scott WJ, Howington J, Feigenberg S, Movsas B, and Pisters K (2007). Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines. *Chest* **132**(3_Suppl.), 234S–242S.
[3] Hirsch FR, Varella-Garcia M, Bunn Jr PA, Franklin WA, Dziadziuszko R, Thatcher N, Chang A, Parikh P, Pereira JR, and Ciuleanu T, et al (2006). Molecular predictors of outcome with gefitinib in a phase III placebo-controlled study in advanced non-small-cell lung cancer. *J Clin Oncol* **24**(31), 5034–5042.
[4] Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, and Christian MC (2000). New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* **92**(3), 205–216.
[5] Win T, Miles KA, Janes SM, Ganeshan B, Shastry M, Endozo R, Meagher M, Shortman RI, Wan S, and Kayani I, et al (2013). Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non-small cell lung cancer. *Clin Cancer Res* **19**(13), 3591–3599.
[6] Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, Liang C, Tian J, and Liang C (2016). Radiomics Signature: A Potential Biomarker for the Prediction of

Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. *Radiology* **281**(3), 947–957.

[7] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, Lambin P, Haibe-Kains B, Mak RH, and Aerts HJ (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* **114**(3), 345–350.

[8] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, and Rietveld D, et al (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* **5**, 4006.

[9] Gillies RJ, Kinahan PE, and Hricak H (2015). Radiomics: images are more than pictures, they are data. *Radiology* **278**(2), 563–577.

[10] Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, and Miles KA (2013). Non–small cell lung cancer: histopathologic correlates for texture parameters at CT. *Radiology* **266**(1), 326–336.

[11] Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O, Hawkins S, Kim J, Goldgof DB, Hall LO, and Gatenby RA, et al (2014). Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Transl Oncol* **7**(1), 72–87.

[12] Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, and Schwartz LH (2016). Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep* **6**, 23428.

[13] Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia AL, Stringfield O, Ye Z, and Gillies RJ (2016). Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. *Clin Lung Cancer* **17**(5), 441–448. e6.

[14] Coroller TP, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee SW, Mak RH, and Aerts HJ (2016). Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol* **119**(3), 480–486.

[15] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, and Pringle M (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* **26**(6), 1045–1057.

[16] Grove O, Berglund AE, Schabath MB, Aerts HJ, Dekker A, Wang H, Velazquez ER, Lambin P, Gu Y, and Balagurunathan Y, et al (2015). Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS One* **10**(3), e0118261.

[17] Song J, Yang C, Fan L, Wang K, Yang F, Liu S, and Tian J (2016). Lung lesion extraction using a toboggan based growing automatic segmentation approach. *IEEE Trans Med Imaging* **35**(1), 337–353.

[18] Tian J, Xue J, Dai Y, Chen J, and Zheng J (2008). A novel software platform for medical image processing and analyzing. *IEEE Trans Inf Technol Biomed* **12**(6), 800–812.

[19] Haralick RM and Shanmugam K (1973). Textural features for image classification. *IEEE Trans Syst Man Cybern* (6), 610–621.

[20] Galloway MM (1975). Texture analysis using gray level run lengths. *Comput Graph Image Process* **4**(2), 172–179.

[21] Baish JW and Jain RK (2000). Fractals and cancer. *Cancer Res* **60**(14), 3683–3688.

[22] Harrell F (2015). Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015 .

[23] Parmar C, Grossmann P, Bussink J, Lambin P, and Aerts HJ (2015). Machine learning methods for quantitative radiomic biomarkers. *Sci Rep* **5**, 13087.

[24] Hoang T, Xu R, Schiller JH, Bonomi P, and Johnson DH (2005). Clinical model to predict survival in chemonaive patients with advanced non–small-cell lung cancer treated with third-generation chemotherapy regimens based on Eastern Cooperative Oncology Group data. *J Clin Oncol* **23**(1), 175–183.

[25] Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, van Baardwijk A, Wanders R, De Ruysscher D, and Steyerberg E (2014). A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. *Radiother Oncol* **112**(1), 37–43.

[26] Cistaro A, Quartuccio N, Mojtahedi A, Fania P, Filosso PL, Campenni A, Ficola U, and Baldari S (2013). Prediction of 2 years-survival in patients with stage I and II non-small cell lung cancer utilizing 18F-FDG PET/CT SUV quantifica. *Radiol Oncol* **47**(3), 219–223.

[27] Sauerbrei W, Boulesteix A-L, and Binder H (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data. *J Biopharm Stat* **21**(6), 1206–1231.

[28] He H and Garcia EA (2009). Learning from imbalanced data. *IEEE Trans Knowl Data Eng* **21**(9), 1263–1284.